Table of Contents

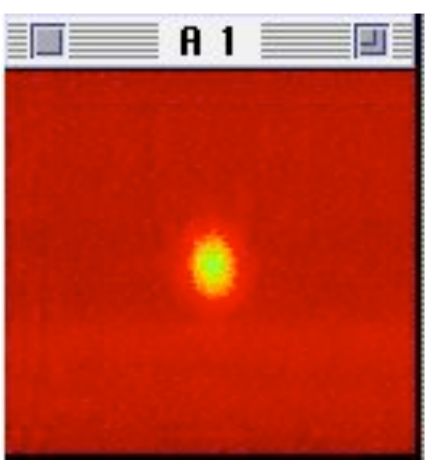
- Motivation
- Simple Statistics--a refresher
  - Probability, means, variances
  - Common Probability Distributions
- Propagation of Errors
  - Analytic & Monte Carlo approaches
- Estimation of Parameters from Data
  - Principle of Maximum Likelihood
  - Development of Least Squares Fitting
    - Linear
    - Non-Linear
  - Chisquare Phenomenology
  - Fitting vs. Simple Statistical Analyisis
- Conclusion (I hope!)

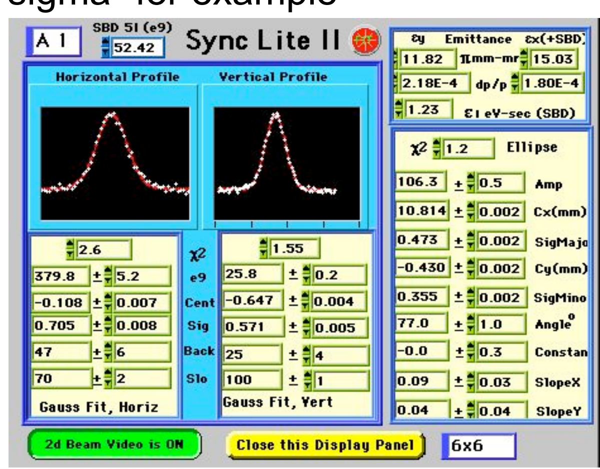
Motivation

- This was a tutorial for the 1994 BIW !!
- At the time I thought it was useful since we were introducing a lot of “Smart” Instruments
  - Parameters supplied to the control room needed to be analyzed from the raw data
    - Flying Wires, Sync Lite, IPM, CPM, SBD
  - Instruments were being developed (hardware and software) by engineers and techs in the department
- Idea was/is to supply a refresher course + some of the backgrounds for how this stuff works.
- And I enjoy this stuff--actually find it fun to do.

Data Reduction and Precision Measurements

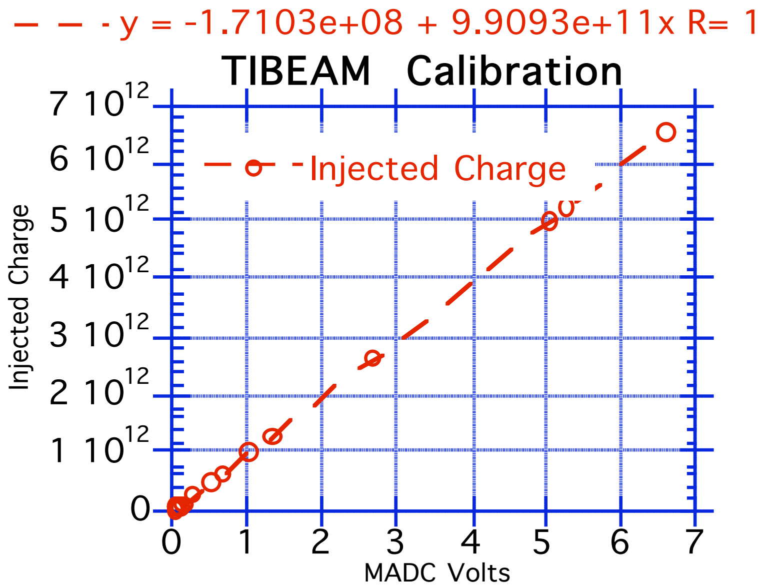
- Some individual instruments produce 100’s of kbytes of data every measurement cycle.
  - Although it is possible to display directly as a comfort display---e.g. Sync Lite Video Beam image, usually the MCR wants a distilled down version of the data--”sigma” for example





Data Reduction and Precision Measurements

- For other instruments, you want to calibrate the detector response



Parameterization of complex data by simple functions

- Sometimes it is useful to apply the Statistical “Mechanics” to extract other useful information
  - For Sync Lite, I wanted to calculate the light produced at the edge of a Tevatron Dipole
    - Fit Magnet factory data to an analytic function that I could then plug into a formula for photon production.

Simple Statistics---a refresher

- It is important when undertaking a measurement that a "True" value actually exists.
- The measurements then represent an attempt to find that "True" value
  - From our data, we often represent the "True " value" from the average, or *mean* of the measurements
  - If we look at the *width* (or spread) of the measured data, then if that *width* is small we have confidence that the *mean* is a good representative of that "True" value
    - A large *width* typically gives us less confidence that we know the "True" value

Simple Statistic---a refresher (2)

- Define $\boldsymbol{P_i}$ for a discrete distribution or $\boldsymbol{P(x)}$ for a continuous distribution as the probability that given the "True" value, that when we make an individual measurement we get the number ***i***, or ***x***
  - Discrete distribution would be like the numbers on a roulette wheel, whereas a continuous distribution would be the probability of measuring a distance.
- We can define the first few moments of the probability distribution

$$\text{Normalization} \quad 1 = \sum_{i=1}^{N} P_i \xrightarrow{continuous} \int_{-\infty}^{\infty} P(x)dx$$

$$\text{Mean} \quad \mu = \sum_{i=1}^{N} iP_i \xrightarrow{continuous} \int_{-\infty}^{\infty} xP(x)dx$$

$$\text{variance} \quad = \sum_{i=1}^{N} i^2 P_i \xrightarrow{continuous} \int_{-\infty}^{\infty} x^2 P(x)dx$$

Simple Statistic---a refresher (3)

- From the variance we can define a *width,*
  - *aka standard deviation* σ
  - aka *sometimes root mean square (rms)*

$$\text{rms} \qquad \sigma = \sqrt{\sum_{i=1}^{N}(i-\mu)^2 P_i} \xrightarrow{\textit{continuous}} \sqrt{\int_{-\infty}^{\infty}(x-\mu)^2 P(x)dx}$$

$$\text{re-writing} \qquad = \sqrt{(\sum_{i=1}^{\infty} i^2 P_i) - \mu^2} \xrightarrow{\textit{continuous}} \sqrt{(\int_{-\infty}^{\infty} x^2 P(x)dx) - \mu^2}$$

$$\text{or} \qquad = \sqrt{\text{var} - \mu^2}$$

Simple Statistic---a refresher (4)

- There are other semi-common moments we use (3rd and 4th) from which we can derive other parameters
  - skew from third moment (symmetry around peak)
    - Does peak=mean?
  - kurtosis or kurtosis "excess" (how "peaky" the distribution is)
    - Gaussian or Normal distribution has kurtosis "excess"=0
- For an arbitrary distribution, all moments matter, but usually we use the low order (1st and 2nd!) moments to simplify discussing the results of a series of measurements

Sample Mean and Variance

- If we make N measurements of a quantity, we can define a *sample mean* and *sample width*
  - Sample mean and sample width are estimates or the probability distribution μ and σ respectively

$$\text{sample mean } \bar{x} = \left(\sum_{i=1}^{N} x_i / N\right)$$

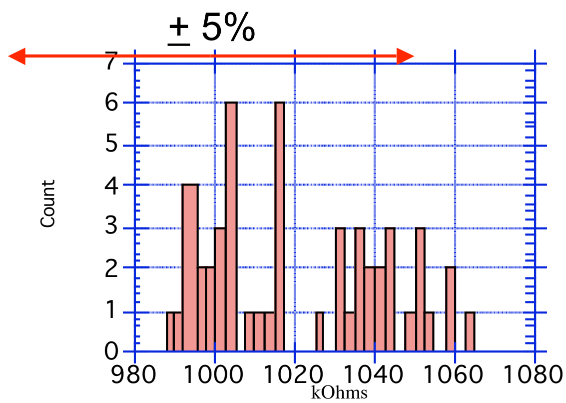
$$\text{sample width } s = \sqrt{\sum_1^N \frac{(x-\mu)^2}{N}} = \sqrt{\overline{x^2} - 2\bar{x}\mu + \mu^2} \xrightarrow{\mu\to\bar{x}} \sqrt{(\overline{x^2} - \bar{x}^2)\left(\frac{N}{N-1}\right)}$$

- This last slight of hand was added to reflect the fact that we have used the data once to calculate s, and in a sense have lost one of the degrees of freedom in the N data points
  - Also note that s is not defined for N=1!
  - In limit of large N, not much difference
- Define $\mu = \lim_{N\to\infty} \bar{x}$ and $\sigma = \lim_{N\to\infty} s$

Sample Mean and Variance:
Resistor measurement

- Measure the resistance of fifty-one 5% 1 MΩ 1/4 watt Carbon resistors
  - All from same box
  - Actually I was interested because I wasn't sure what the 5% spec really meant.
  - Used Fluke DMM
    - Manufacturer spec was 0.5% accuracy for resistance measurements
    - Measured a single resistor many times and found fluctuations were < 0.2%
      - Precision of DMM (and my measuring) was pretty good
    - How to verify absolute accuracy??

Sample Mean and Variance:
Resistor measurement results



- Mean=1021 kΩ
- s = 22kΩ (2.2%)
- Standard deviation of Mean=3.1kΩ
  - not yet defined.
- What is significance of the result?

Common Probability Distributions- Binomial

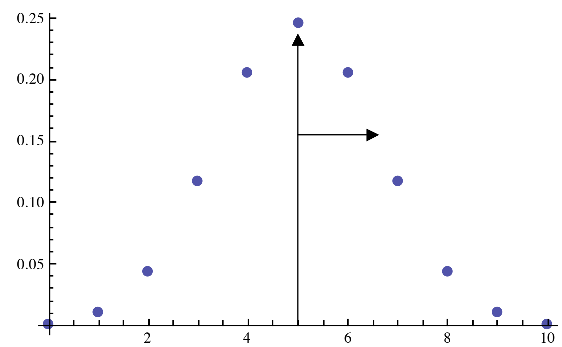
$$P(m,p,N) = \frac{N!\,p^m (1-p)^{N-m}}{m!(N-m)!}$$

- Gives the probability of *m* successful outcomes out of N *independent* trials
  - Probability of success = *p*
    - Probability of failure = *1-p*
- This is a Discrete Distribution since the observables, *m* (# of successes) are integers.

$$\mu = Np \;\text{ and }\; \sigma = \sqrt{Np(1-p)}.$$

Common Probability Distributions- Binomial--Examples

- Flip a coin 10 times
- How many times you get tails?
  - 10 tails 0.00098
  - 9 tails 0.00977
  - 8 tails 0.04395
  - 7 tails 0.11719
  - 6 tails 0.20508
  - 5 tails 0.24609
  - 4 tails 0.20508



- μ = 5.00 (of course!)
- σ = 1.58
- $\sigma/\mu = 32\%$

Common Probability Distributions- Poisson

$$P(m,\mu) = \frac{\mu^m e^{-\mu}}{m!}$$

- Probability of *m* successful observations when the mean is μ
- Poisson distribution can be shown to be a limit of Binomial distribution
  - N (number of trials) >>>1
  - *p* (probability of success) <<<<1
  - However limit N**p* -> *finite* = μ
- One beloved feature of Poisson Distribution
  - $\sigma = \sqrt{\mu}$
- Example is Radioactive decay

Common Probability Distributions-
Normal *aka* Gaussian

$$P(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{(x-\mu)}{\sigma}\right)^2}$$

- Probability of observing *x* when the mean is μ and standard deviation = σ
- Normal distribution can be shown to be a limit of Binomial distribution when
  - N (number of trials) >>>1
- This is a continuous distribution
- The Gaussian or Normal (or Bell Shaped) Distribution is found everywhere, and we will later see it is a limit of many distributions.

Common Probability Distributions-
Uniform Distribution

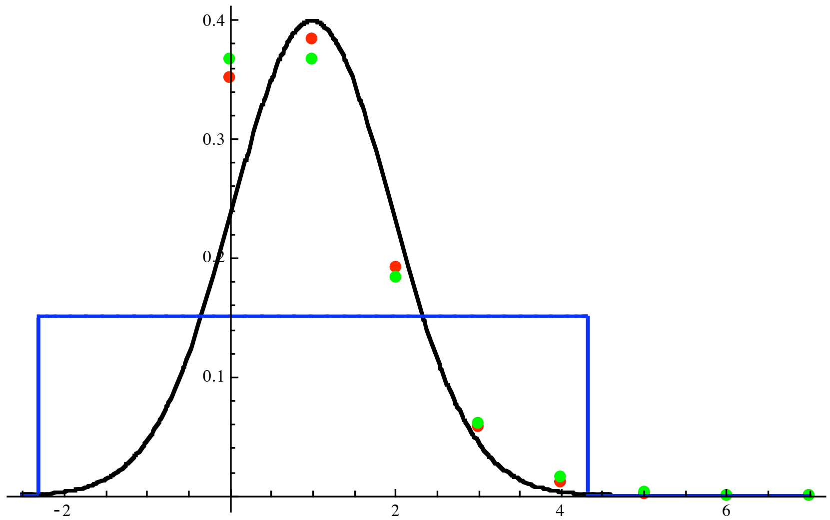
$$P(x,\mu,w) = \frac{1}{w} \text{ if } |x - \mu| \le \frac{w}{2}$$

$$= 0 \text{ if } |x - \mu| \ge \frac{w}{2}$$

- Continuous Distribution
- Easily generated on calculators by scaling the random number generator built in
- $\sigma = \sqrt{w^2/12}$
- Can often serve (with care of course) as a “poor man’s” Gaussian using μ and σ accordingly.

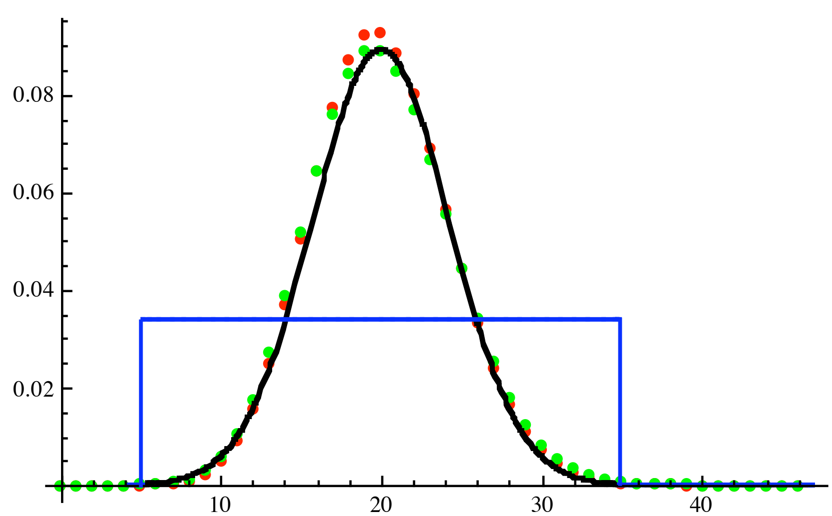
Common Probability Distributions Comparison

- Binomial, Poisson, Gaussian, Uniform
- 12 trials, p=1/12, others are scaled to match μ & σ
  - $\mu = 1$, $\sigma \sim 1$
  - Note first two have points only for x>1



Common Probability Distributions Comparison

- Binomial, Poisson, Gaussian, Uniform
  - 240 trials, p = 1/12, others are scaled to match μ and σ
  - $\mu = 20$, $\sigma = 4.3$



Propagation of Errors-Analytic Approach

- Suppose we want to measure the length "L" and width "W" of a rectangle and want to determine the area "A".
  - There are uncertainties in "L" and "W", (σ_L and σ_W)
  - What is the uncertainty in "A", σ_A?

$$A = L * W$$

$$A - A_t = W * (L - L_t) + L * (W - W_t)$$ Expand about "true" values

$$\text{or} \qquad dA = WdL + LdW$$

$$\overline{dA^2} = \overline{(WdL)^2} + \overline{(LdW)^2} + \overline{LWdLdW}$$

$$\sigma_A = \sqrt{(W\sigma_L)^2 + (L\sigma_W)^2}$$

Square and average, since L and W measurements are considered independent Of each other, cross terms average to 0.

Propagation of Errors-Analytic Approach

- In general for a quantity R(r_1, r_2, r_3,…r_i …)

$$\sigma_R^2 = \left\langle \left(\sum_i \frac{\partial R}{\partial r_i} dr_i \right)^2 \right\rangle \Rightarrow \sum_i \left\langle \left(\frac{\partial R}{\partial r_i} d_{r_i} \right)^2 \right\rangle = \sum_i \left(\frac{\partial R}{\partial r_i} \sigma_{r_i} \right)^2$$

- Some common functions

$$R = r_1 + r_2, \qquad \sigma_R = \sqrt{\sigma_{r_1}^2 + \sigma_{r_2}^2}$$

$$R = r_1 * r_2, \qquad \sigma_R / R = \sqrt{(\sigma_{r_1}/\mathrm{r}_1)^2 + (\sigma_{r_2}/\mathrm{r}_2)^2}$$

$$R = \frac{r_1}{r_2}, \qquad \sigma_R / R = \sqrt{(\sigma_{r_1}/\mathrm{r}_1)^2 + (\sigma_{r_2}/\mathrm{r}_2)^2}$$

$$R = r_1 r_2^2, \qquad \sigma_R / R = \sqrt{(\sigma_{r_1}/\mathrm{r}_1)^2 + (2\sigma_{r_2}/\mathrm{r}_2)^2}$$

$$R = \cos(r), \qquad \sigma_R / R = |\tan(r)|\sigma_r$$

Propagation of Errors-Analytic Approach

Example- Uncertainty in the mean determined from N measurements

- Recalling how we make the mean (average) from a quantity of measurements, you can see that we should be able to determine the uncertainty in the mean by propagating the error using the formulism just developed

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \text{ and } d\bar{x} = \frac{1}{N}\sum_{i=1}^{N} dx_i, \text{ giving}$$

$$\sigma_{\bar{x}} = \frac{\sqrt{\sum_{i=1}^{N}\sigma_{x_i}^2}}{N}, \text{ if } \sigma_{x_i} \text{ are the same} = \sigma_x, \text{ then}$$

$$\sigma_{\bar{x}} = \frac{\sigma_x}{N}\sqrt{\sum_{i=1}^{N} 1} = \frac{\sigma_x}{N}\sqrt{N} = \frac{\sigma_x}{\sqrt{N}}$$

- This is a very important derivation!
  - For example, doubling the number of measurements only will improve the statistical uncertainty in the mean by ~41%

Example of Uncertainty in Mean

- Suppose we toss a coin N=100 times, and record the number of times it comes up tails.
  - Call this X a “Run”
  - Now lets do M Runs like that.
- Call $P_{est} = \mu$ of the M Runs of X
  - How big should M be if we want to measure P_{est} to 1%?
  - Flipping a coin is covered by the Binomial distribution
    - Each Run of N=100 tosses of X should be distributed with a mean of Np = 50 and $\sigma = (Np(1-p))^{1/2} = 5$.
    - 1% uncertainty in 50= 0.5, so to reduce 5 to 0.5
      - $\sigma_\mu = \sigma/M^{1/2} = 5/M^{1/2} < 0.5$ or M~100
      - Equivalent to 10000 flips (MxN)! (which is another way to look at problem)
- See LV demo.

Propagating Errors- Monte Carlo Technique

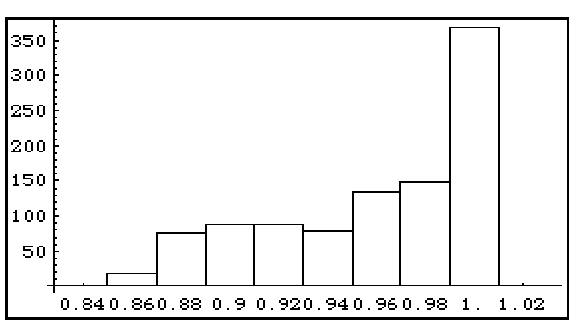
- Analytic approach to error propagation basically assumed that only the first order terms in a Taylor's expansion are important.
  - There is a problem when R is at a (functional) maximum or minimum, because then the first derivatives are =0
    - Note R=cos(r) when r is near 0.

$$R = \cos(r), \quad \sigma_R / R = |\tan(r)|\sigma_r$$

    - Problem is that in this case you need to consider the higher order derivatives.

Propagating Errors- Monte Carlo Technique

- Sometimes it is just easier to consider a Monte Carlo technique to understand how a quantity depends on its many variables.
- Randomly generate the independent variable with an appropiate width distribution that reflects actual distribution.
  - For each set of randomly generated variables r_i, calculate R.
  - Histogram R to see its distribution
- Histogram of R = cos(r) for r generated from uniform distribution with r = 0.0$\pm$0.1
  - Note function is not symmetric around cos(r)=1



Cos(r)

Central Limit Theorem-

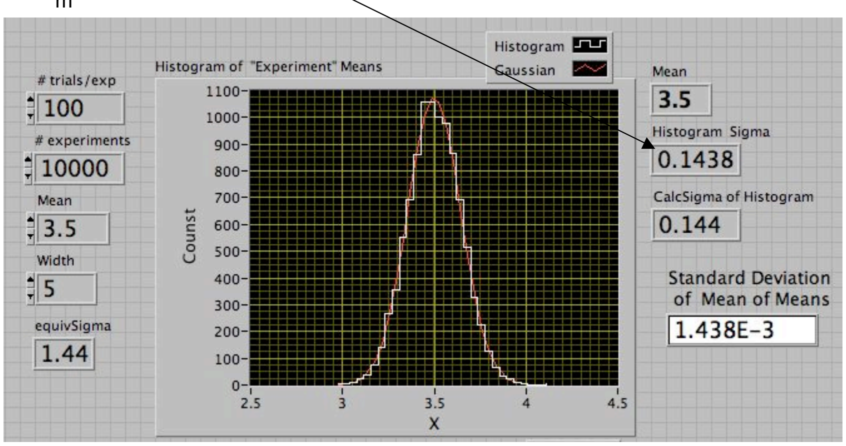
Why Daddy is everything a Gaussian?

- Given an arbitrary distribution which has the mean μ and variance (σ^2) defined:
- Central Limit theorem says that the the Average of the N measurements is distributed as a Gaussian with standard deviation $\sigma/N^{1/2}$
- Handwaving explanation is that most of our measurements are really due to the average of many processes at the microscopic level.
  – Example: DMM Current measurements are measuring $\sim 10^{23}$ electrons, each moving according to a Maxwell Boltzman Distribution. Yet our measurement will look (probably) gaussian.

Central Limit Theorem-

Illustration of Central Limit Theorem

- The Histogram of 10000 “experiments” (LV Demo)
  – Each experiment involves calculating the mean value for N= 100 data points generated from a Uniform Distribution of Width=5units and Mean=3.5units (σ=1.44)
  – Histogram in White is of the means from each experiment
  – Red curve is a Gaussian whose area=10000 with $\sigma_m = \sigma/N^{1/2} = 0.1^* \sigma$



Estimation of Parameters from Data

- Up to now we tend to have been talking about measuring a single quantity "R_i" at a single point "X" (or perhaps at a common 2 or 3 D point (x,y,z). Sometimes x is just the index.
- Now lets consider that we measure a function f(x) at multiple x points.
- We assume we *a priori* know the functional relationship between f and x
- E.g.

$$\text{polynomial} \quad \mathrm{f(x)} = \mathrm{a}_0 + \mathrm{a}_1 x + \mathrm{a}_2 x^2 \text{ or}$$

$$\text{gaussian} \quad \mathrm{f(x)} = \mathrm{a}_0 e^{-\frac{1}{2}\left(\frac{x-\mathrm{a}_1}{\mathrm{a}_2}\right)^2} + \mathrm{a}_3 x + \mathrm{a}_4$$

Estimation of Parameters from Data

- What we measure are the f_i values at the points x_i, and we would like to find the parameters a0, a1, a2,… which describe our data the best.
- How do we do this??

Principle of Maximum Likelihood

- aka ***Principle that nature plays fair!***
- The Principle of Maximum Likelihood says the values of the parameters a_0, a_1, a_2,…which maximizes the probability of measuring our data points $f(x_i)$ are the best estimates we can have of those parameters.

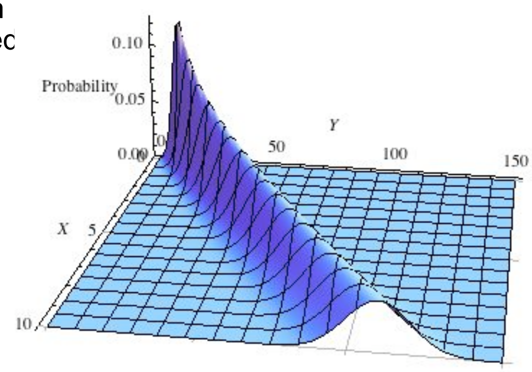
End Pt 1

Principle of Maximum Likelihood-Example

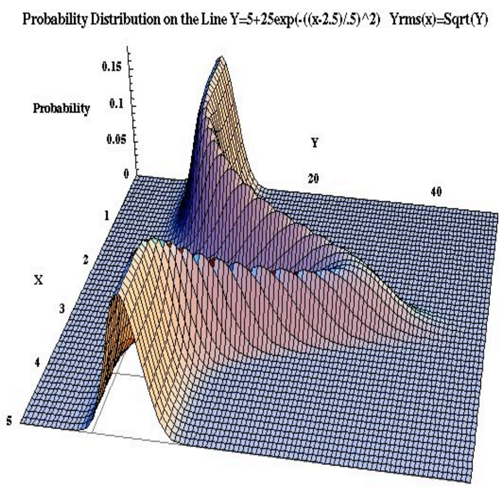
- Suppose we measure a series of points y_i at x_i. $(y(x_i))$
  - Furthermore lets say we believe that a linear relationship exists between y and x
    - $y(x_i)=a_0+a_1x$
    - For the sake of argument, we will assume that there is no uncertainty in the "x" value, but all measurement uncertainties are in y. In other words we pick an "x" position, and measure y.
    - The uncertainty ("error") in each y_i measurement is described by a probability distribution ----the uncertainties may vary at different y_i values.
      - $Y_{i,}\ \sigma_{i,}\ x_i$
        - » Note the uncertainty may involve more than just a single σ_I, but we will label it that way here.

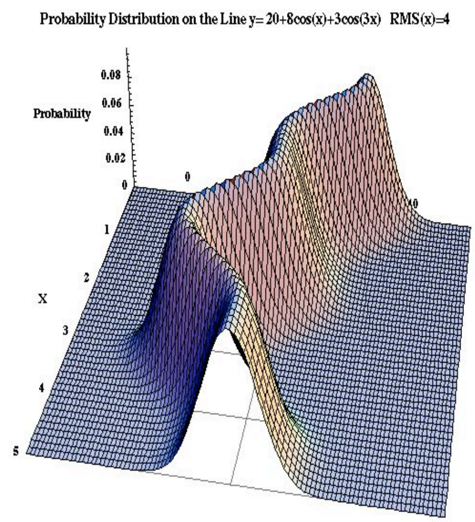
Principle of Maximum Likelihood-Example

- Probability of measuring y_i at x_i
  - The crest of the curve represents the "true" curve that generates our particular data points.
    - We can measure a "y" at each point "x" , the relative probabilty being given by the curve.
    - You can see it is most probably, in this case to measure a point on the crest
    - The probability distribution shown here is a normalizec Gaussian, whose σ varies as $y^{1/2}(x)$.
    - The crest function is $y(x)=10+10x$



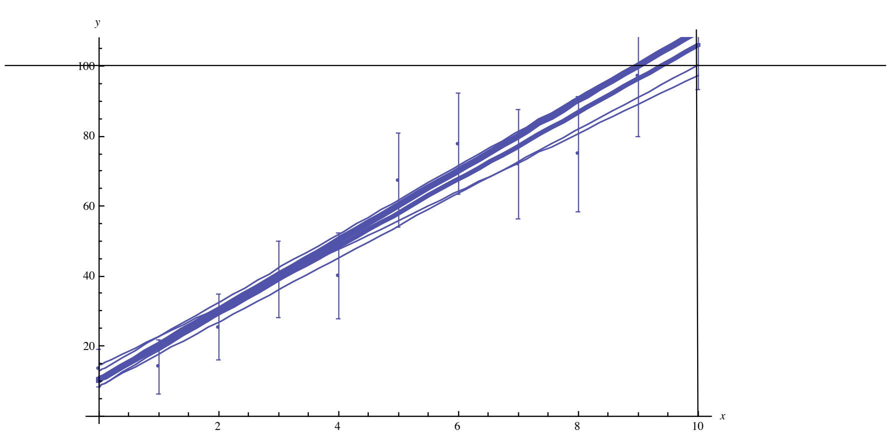
Some other examples





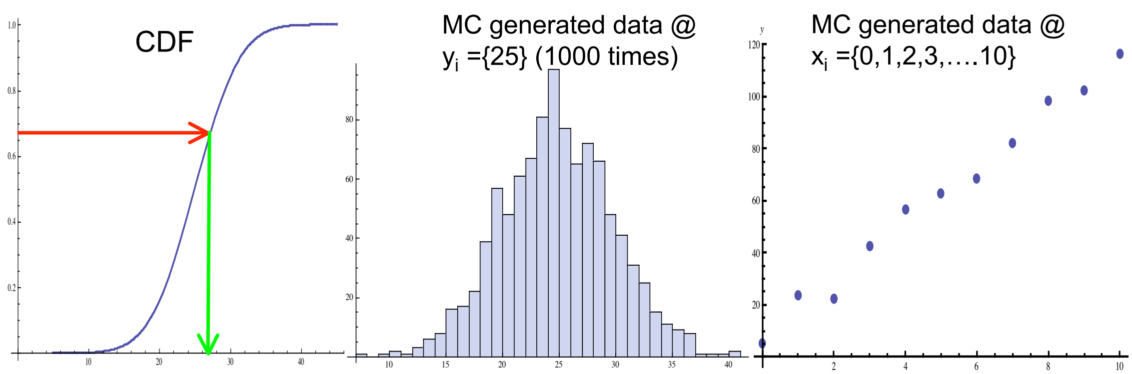
Principle of Maximum Likelihood-Example

- We will generate some data using a Monte Carlo generator, for
  - Uniform distribution
    - y_i= $(10+10*x_i)*[w_{i*}(ran-0.5)]$
      - "ran" generates a number between 0-1
      - with w_i= $y_i^{1/2}$
    - The error bars shown on the data are the uniform widths.
    - The explanation for the curves will be explained later



Principle of Maximum Likelihood-Example

- Generate data using a Normal uncertainty distribution with
  - $\sigma_y = y^{1/2}$
- This is a Cumulative Distribution Function for a Normal Distribution
  - Just the integral of the Normal distribution function
  - centered at y_i=25 & $\sigma_{yi} = 25^{1/2}$ =5
  - Use “ran” to generate a number between 0->1, say get 0.67
  - Find 0.67 on vertical axis, find value for “yi”(~26.9) on horizontal axis
    - histogram uses procedure to generate 1000 “datapoints”
    - RHS Plot shows “data” y_i for x_i ={0,1,2,3,….10} (y(x)=10+10x)

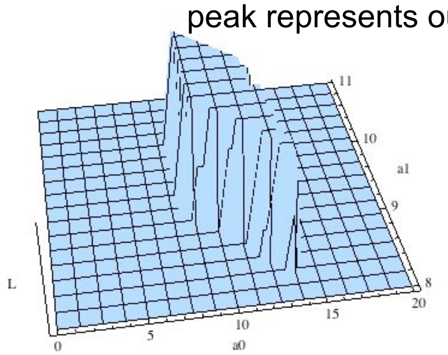


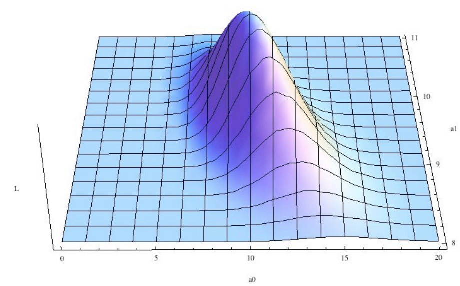
Maximum Likelihood- How to calculate?

- We have the data points {y_i} and the probability distributions for each point
  - Recall Uniform and Normal Distributions which scale their widths (w and σ) as Sqrt[y_i]
- Probability to see the particular set of data points is just
  - Ptotal = $\Pi_i P_i$
  - Where P_i=Probability to observe y_i given $y(x_i,a_0,a_1)$
    - Scan over {a_0,a_1}
      - For each set use $y(x_i,a_0,a_1)$ as the hypothesis, used to make probability distribution. Then calculate the probability to measure measure y_i for that distribution.
      - Multiply all individual Pi together,…
      - Go to next {a_0,a_1} set and redo.

ML-Examples

- Left graphic is Uniform ML, Right is Normal ML
  - Note that Uniform Distribution is either 0 or a max value, because the probability for any data point is either 1/w if within window, or =0.
    - The lines shown on 2D plot were taken from the edges of this parallelogram
    - What is best estimate? We cannot tell using on ML method
      - Maybe we could choose the center or the parallelogram, but is is no more likely than any other.
  - For Normal distribution, there is a peak in a0 and a1. This peak represents our best estimate of the parameters





Development of Least Squares Fitting.

- If the underlying Probability distributions are "Normal" Distributions, then some wonderful things happen.
- In this case the Likelihood function is

$$L = \prod_i^n P\left(y_i, y(x), \sigma(x_i)\right) = \prod_i^n \frac{1}{\sqrt{2\pi}\sigma(x_i)} e^{-\frac{1}{2}\left(\frac{(y_i - y(x_i))}{\sigma(x_i)}\right)^2}$$

$$= \left(\prod_i^n \left(\frac{1}{\sqrt{2\pi}\sigma(x_i)}\right)\right) e^{-\frac{1}{2}\sum_i^n \left(\frac{(y_i - y(x_i))}{\sigma(x_i)}\right)^2}$$

Development of Least Squares Fitting (2).

$$\left(\prod_i^n\left(\frac{1}{\sqrt{2\pi}\sigma(x_i)}\right)\right)e^{-\frac{1}{2}\sum_i^n\left(\frac{(y_i-y(x_i))}{\sigma(x_i)}\right)^2}$$

- y_i is our data
- $y(x_i$, a0,a1,a2,…..) is our hypothesis
- $\sigma(x_i)$ is our estimate of the uncertainty (may depend on x_i but in principle not on the a_i)
  - If last statement is true, then the complete dependence of the likelihood depends on the argument of the exponential
- Maximizing L is equivalent to minimizing
- Chisquare $\chi^2 = \sum_i^n\left(\frac{(y_i - y(x_i))}{\sigma(x_i)}\right)^2$

Linear Least Squares Fit

- We can gain some insight by expanding χ^2 in a second order Taylor's series about its minimum

$$\chi^2(a) = \chi^2(a^0) + \sum_j \frac{\partial\chi^2(a^0)}{\partial a_j}da_j + \frac{1}{2}\sum_{j,k}\frac{\partial^2\chi^2(a^0)}{\partial a_j\partial a_k}da_j da_k$$

- First term is just value of χ^2 at the minimum
- Since we are at the minimum the second term

$$\frac{\partial\chi^2(a^0)}{\partial a_j} = \frac{\partial\sum_i\left(\frac{y(x_i,a)-y_i}{\sigma_i}\right)^2}{\partial a_j} = 2\sum_i\left(\frac{y(x_i,a)-y_i}{\sigma_i^2}\right)\frac{\partial y(x_i,a)}{a_j} = 0$$

- Nomenclature is "i" is the "ith" data point, "j" is the "jth" parameter a_j

Linear Least Squares Fit (2)

- To be = 0 for arbitrary da_j, each of the linear terms is set =0
- IF $y(x_i,a_0,a_1,\ldots)$ is a LINEAR function of the a's

$$y(x_i,a) = \sum_j f_j(x_i) a_j$$

- The math simplifies even more, and we have what is known as a Linear Least Squares Fit
  - Note that y(x,a) is not necessarily linear in x, the independent variable
  - Doing the derivative and collecting terms we end up with a matrix equation
  - $\{\alpha\}a=\beta$,
    - $\{\alpha\}$being known as the "curvature" matrix, "a" is just the parameter vector and β the data vector

Linear Least Squares Fit (3)

in component form $\sum_k \alpha_{jk} a_k - \beta_j = 0$, and

$$\alpha_{jk} = \sum_i \frac{f_j(x_i) f_k(x_i)}{\sigma_i^2}, \quad \text{and } \beta_j = \sum_i \frac{y_i f_j(x_i)}{\sigma_i^2}.$$

- Note that $\{\alpha\}$ does not actually depend on the measured data!
  - Does depend inversely on the square of the uncertainties
- a_k are just our desired parameters that we want to determine
- The β vector holds all the dependence on our actual measured data (in the y_i)
- Solution means we need to invert $\{\alpha\}\text{-> }\{\alpha\}^{-1}=\{\varepsilon\}$

LLSF (4)

$$(a) = \{\alpha\}^{-1}(\beta) = \{\varepsilon\}(\beta) \text{ or}$$

$$a_j = \varepsilon_{jk}\beta_k \quad .$$

- $\{\varepsilon\}$ is also known as the "Error" Matrix
  - Can probably guess what that is going to mean!

- 2nd Derivative terms from χ^2

$$\frac{1}{2}\sum_{j,k}\frac{\partial^2\chi^2\left(a^0\right)}{\partial a_j \partial a_k} da_j da_k$$

- it contains $\{\alpha\}$, the "curvature" matrix
  - Now you can see why we called it that
  - $\{\alpha\}$ represents how curved the surface is
    - Recall that $\{\alpha\}$ was inversely proportional to σ_i^2, the uncertainties in the measurements. Small σ_i implies a steeply rising χ^2 surface.

LLSF (5)- Errors of the determined parameters a_i

- Now that we can determine the "fit",it is reasonable to ask how well are the parameters known.
- Since the source of uncertainty was the data itself, the error in the a_i must come from there.

$$da_j = \sum_i \frac{\partial a_j}{\partial y_i} dy_i = \sum_i \frac{\partial\left(\sum_k \varepsilon_{jk}\beta_k\right)}{\partial y_i} dy_i = \sum_k \varepsilon_{jk} \sum_i \frac{\partial \beta_k}{\partial y_i} dy_i$$

…Lots of formulae…<snip> …see preprint

$$\sigma_{jl}^2 = \left\langle da_j da_l \right\rangle = \sum_{k,m} \varepsilon_{jk}\varepsilon_{lm}\alpha_{km} = \sum_k \varepsilon_{jk}\delta_{lk} = \varepsilon_{jl}$$

- Or as was hinted, the elements ε_{jl} of $\{\varepsilon\}$, the error matrix are related to the errors in the parameters
  - Note that in general the non-diagonal terms are not zero

LLSF:Example

- Use y(x)=10+10x, generate

$$x_i = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]$$
$$y_i = [13.6, 14.1, 25.3, 38.9, 40.0, 67.3, 77.8, 72.0,75.0, 97.1, 115.3]$$
$$\sigma_{y_i} = [3.2, 4.5, 5.5, 6.3, 7.1, 7.8, 8.9, 9.5, 10.0, 10.5]$$

$$\{\alpha\} = \begin{Bmatrix} \sum_{i=1}^{11} \frac{1}{\sigma_{y_i}^{\ 2}} & \sum_{i=1}^{11} \frac{x_i}{\sigma_{y_i}^{\ 2}} \\ \sum_{i=1}^{11} \frac{x_i}{\sigma_{y_i}^{\ 2}} & \sum_{i=1}^{11} \frac{x_i^2}{\sigma_{y_i}^{\ 2}} \end{Bmatrix} = \begin{Bmatrix} 0.302 & 0.798 \\ 0.798 & 4.702 \end{Bmatrix}$$

$$(\beta) = \begin{pmatrix} \sum_{i=1}^{11} \frac{y_i}{\sigma_{y_i}^{\ 2}} \\ \sum_{i=1}^{11} \frac{y_i x_i}{\sigma_{y_i}^{\ 2}} \end{pmatrix} = \begin{pmatrix} 10.67 \\ 53.00 \end{pmatrix} \qquad \{\varepsilon\} = \{\alpha\}^{-1} = \begin{Bmatrix} 6.00 & -1.02 \\ -1.02 & 0.38 \end{Bmatrix}.$$

LLSF :Example (2)

- Solving for the parameters and errors

$$(a) = \{\varepsilon\}(\beta) = \begin{pmatrix} 10.1 \\ 9.56 \end{pmatrix} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \text{constant} \\ \text{slope} \end{pmatrix} \qquad \sigma_a = \begin{pmatrix} \sqrt{\varepsilon_{11}} \\ \sqrt{\varepsilon_{22}} \end{pmatrix} = \begin{pmatrix} 2.4 \\ 0.62 \end{pmatrix}$$

- Recall data generated with a_0, a_1=10

- Also $\chi^2 = 10.4/9$ degrees of freedom
  - Degrees of freedom = # data points-number of fit parameters (=11-2 in this case)
  - What does value χ^2 mean?

$$\chi^2 = \sum_i^n \left(\frac{(y_i - y(x_i))}{\sigma(x_i)} \right)^2$$

If we have estimated the errors correctly, (data-theory)/error ~1 per term (for a normal distribution), so a “good” fit would give a χ^2 ~ #data points - #times data was used.

LLSF--Using Fit to interpolate

- Assume the last example was a calibration of a voltage as a function of an ADC reading.
- For an arbitrary ADC reading (within the bounds of our fit), how good is the calibration?
  - = as good as the original data points nearby?
  - = better?
- Hopefully better, since we have used all the data to get the fit, so we should do better than any individual measurment
- Will use propagation of errors to find uncertainly of $y(x,a_0,a_1,a_2,\ldots)$

LLSF--Using Fit to interpolate (2)

$$y(x,a_0,a_1) = a_0 + a_1 x$$

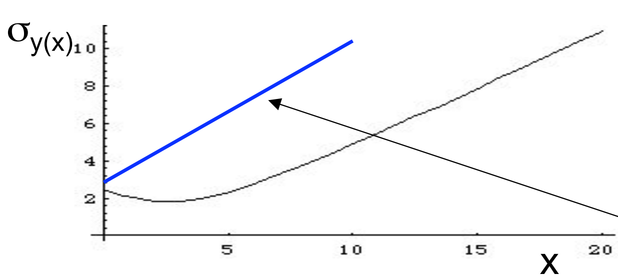
$$dy(x,a_0,a_1) = \sum_j \frac{\partial y(x,a)}{\partial a_j} da_j = 1da_0 + xda_1$$

$$\sigma^2_{y(x,a_0,a_1)} = dy(x,a_0,a_1)^2 = \left(da_0 + xda_1\right)^2 = \sigma^2_{a_0} + x^2\sigma^2_{a1} + 2x\sigma^2_{a_0a_1}$$

$$\sigma_{y(x,a_0,a_1)} = \sqrt{\varepsilon_{00} + x^2\varepsilon_{11} + 2x\varepsilon_{01}}$$

$$\{\varepsilon\} = \{\alpha\}^{-1} = \begin{Bmatrix} 6.00 & -1.02 \\ -1.02 & 0.38 \end{Bmatrix} \text{ see previous slide}$$

$$\sigma_{y(x,a_0,a_1)} = \sqrt{6.00 + 0.38x^2 - 2.04x}$$



- Recall that we fit between $0 \le x \le 10$
- Within this range is interpolation
- Outside is extrapolation
- Not how error grows for x outside this range!
- Our interpolating error ranges from ~2 (@x=~3) to ~5 @x=10
- y(x)=10+10x = original generator

Errors in original data

- (Sqrt(y), or ~3 at x=0 to ~11 @x=10

Non-Linear LSF- Log and other end-runs

$$y = a_1 e^{a_2 x} \quad \rightarrow \log(y) = \log(a_1) + a_2 x \text{ and}$$
$$\text{y} = a_1 x^{a_2} \quad \rightarrow \log(y) = \log(a_1) + a_2 \log(x).$$

- Note the parameters a_i do NOT appear linearly
  - Appears our formulism fails!
- However we can transform the original equations into a new form where new parameters ($\log(a_1)$ and a_2) do appear in a linear manner
  - Caveat: the data point errors are also transformed and need to be handled carefully.
    - If nothing is done-typical, then the low values of y will be overweighted in fit .

Non-Linear LSF- Linearization

- Well since we know how to do linear fits, we will expand $y(x,a_j)$ with respect to a_j

$$\text{Let } da_j = a_j - a_j^0, \text{with } a_j^0 = constant$$

$$y(x,a) = y(x,a^0) + \sum_j \frac{\partial y(x,a^0)}{\partial a_j} da_j, \text{ and}$$

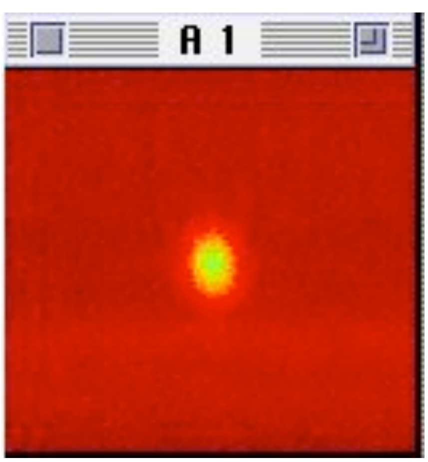
$$\chi^2 = \sum_i \frac{\left(y(x_i,a) - y_i\right)^2}{\sigma_i^2} \rightarrow \sum_i \frac{\left(y(x_i,a^0) - y_i + \sum_j \frac{\partial y(x,a^0)}{\partial a_j} da_j\right)^2}{\sigma_i^2}.$$

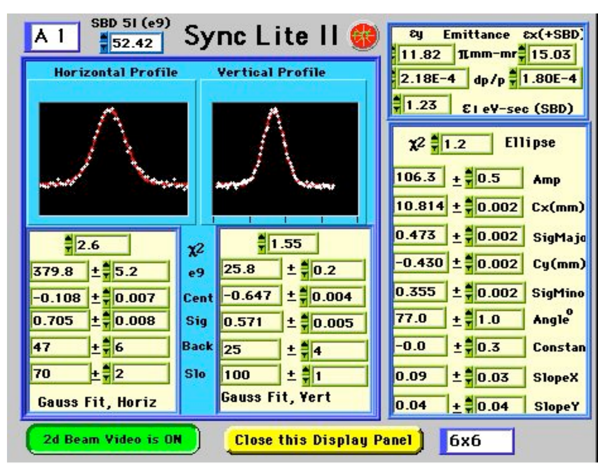
- Note that y(x,a) is a linear function of the "da_j" which is what we will use the L-LSF mechanics to solve.

Non-Linear LSF- Linearization (2)

- How it works.
  - You need to supply the original values of a^0_j
    - This is the hard work.
  - Once this is done, then the $\{\alpha\}$ matrix and β vectors can be calculated (same nomenclature as before)
  - $\{\alpha\}$ is inverted to make $\{\varepsilon\}$,
    - The linear parameters vector **da**= $\{\varepsilon\}\beta$
    - We make $\mathbf{a^1=a^0+da}$, the new estimate for the parameters
      - This new $\mathbf{a^1}$ is plugged back where $\mathbf{a^0}$ was used before, and we calculate new **da**= $\{\varepsilon\}\beta$ and so on.
      - Continue iteration until some condition is satisfied.
        - » Maybe χ^2 < some limit
        - » Maybe **da** $<<\sigma_{\mathbf{a}}$ (the errors in the parameters)--my favorite

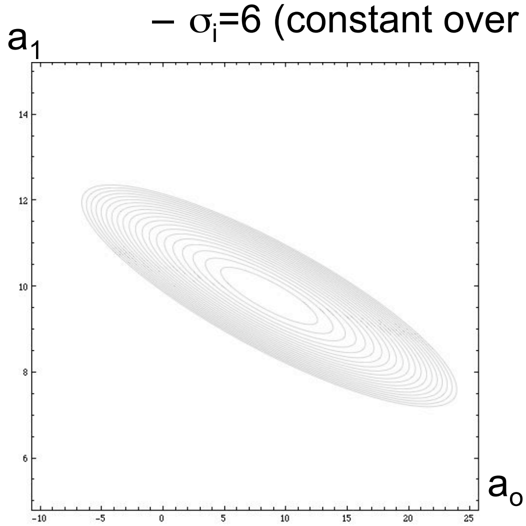
Example (reprise)





Chisquare Phenomenology

- Lets make a 2D Contour map of chiquare for our favorite LSF to $y(x)=a_0+a_1x$
  - Using data generated from $y(x)=10+10x$
  - $\sigma_i=6$ (constant over all x)



contour step size is one unit of χ^2

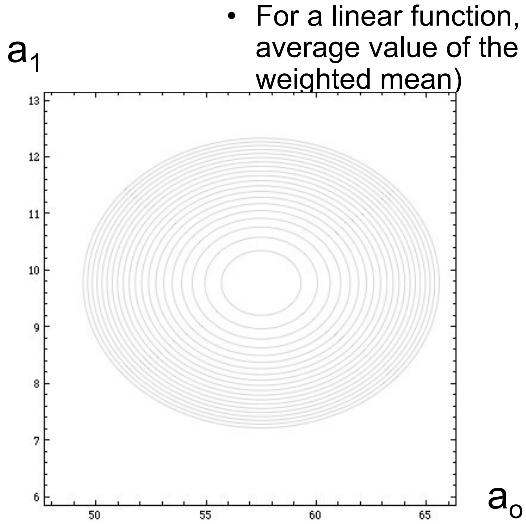
Curvature matrix is

$$\{\alpha\} = \begin{Bmatrix} 0.302 & 0.798 \\ 0.798 & 4.702 \end{Bmatrix}$$

The ellipse is rotated because the curvature matrix is not diagonal

Chisquare Phenomenology (2)

- Same data as before $\sigma_i=6$=constant
  - But fit to $y(x)=b_0+b_1(x-5)$
  - Note we still have 2 parameters, but the functions are different.
  - When x=5, $y(x)=b_{0.}$
    - For a linear function, the center of the x range will be the average value of the function (for equal errors,otherwise use weighted mean)



contour step size is one unit of χ^2

Curvature matrix is now

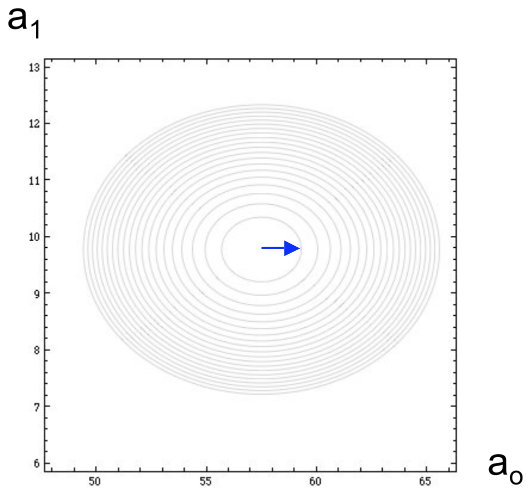
$$\{\alpha\} = \begin{Bmatrix} 0.306 & 0 \\ 0 & 3.06 \end{Bmatrix}$$

The ellipse is rotated because the curvature matrix is not diagonal

Chisquare Phenomenology (3)

- Kind of cool, but no one really does this, but instead I wanted to make a point about the χ^2 contours
- Lets see what happens to χ^2 when we increase a_0 by its error σ_{a0} starting from the minimum central value

$$\chi^2\left(a_0^{best}+\sigma_{a_0}\right)= \chi^2\left(a_0^{best}\right)+\alpha_{00}\sigma_{a_0}{}^2=\chi^2\left(a^0\right)+\alpha_{00}\varepsilon_{00}=\chi^2\left(a^0\right)+1$$



The last step occurs because ε_{00} is the inverse of α_{00}.
So conclusion, is that a the

This is also true in previous rotated ellipse, but you need to re-optimize the other parameters after step

Simple Statistical analysis instead of fitting

- Sometimes it seems desirable not to try and fit data, but instead calculate μ and σ by using standard mean and rms calculations
  - Reason is that LS fitting, especially non-linear is prone to diverging
    - use to mean computer crashes back in the “good old days”
  - Simple μ and σ always give an answer
    - Sometimes not a good answer tho’!

Simple Statistical analysis instead of fitting (2)

- Consider a histogram of data $h(x_i)$
  - Can define the mean and σ by

$$\mu = \frac{\sum h(x_i)x_i}{\sum h(x_i)} \text{ and } \sigma = \sqrt{\frac{\sum h(x_i)x_i^{\ 2}}{\sum h(x_i)} - \mu^2}$$

  - Issues
    - What about background
      - If flat, it doesn't shift μ, but does dilute its statisical significance
        - » However completely screws up the σ calculation
      - Need to carefully subtract it, especially if it is sloping
    - What about (lousy) statistics
      - Really can be an issue with a large fluctuation at large x_i for σ

Simple Statistical analysis instead of fitting (3)

- See LV Demo

Conclusion

- This wasn't an exhaustive course on statistics
  - Maybe exhausting
- Hopefully you may come away with a better appreciation on what the underpinnings of all those neat software packages.